A Unit Selection Text-to-Speech Synthesis System Optimized for Use with Screen Readers

Aimilios Chalamandaris, Sotiris Karabetsos, *Member*, IEEE, Pirros Tsiakoulis, *Member*, IEEE, and Spyros Raptis, *Member*, IEEE

Abstract — *Currently, unit-selection text-to-speech technology is the common approach for near-natural speech synthesis systems. Such systems provide an important aid for blind or partially-sighted people, when combined with screen reading software. However, although the overall quality of the synthetic speech achieved by such systems can be quite high, this fact alone does not guarantee a high level of user satisfaction. Many issues have to be coped with in order to fulfill users' expectations when integrating such systems with screen reading tools aiming to assist blind users. This work describes the design and the implementation approaches for the efficient integration of this technology into screen reading environments. In particular, the issues of natural language processing, speed optimization, multilingual design and overall quality optimization are mainly addressed in this paper. In order to evaluate the resulting system, we carried out subjective assessment tests where expert users provided feedback about performance, quality and overall experience.*[1]

Index Terms —Speech Synthesis, Unit Selection, Text-to-Speech, Screen Reader, Assistive Technology.

I. INTRODUCTION

Text-to-Speech (TTS) technology aims to produce synthetic voice from textual information, thus serving as a more natural interface in human machine interaction. Nowadays that near-natural synthetic speech has been achieved, TTS systems are widely adopted in everyday solutions. Following a significant research progress in the field of speech synthesis, the unit selection concatenative method has become the dominant approach for building naturally sounding text-to-speech systems. This technique relies on the runtime selection and compilation of speech units from a large speech database [1]. The speech database usually derives from a sufficiently large corpus where appropriately selected spoken utterances are carefully annotated to the unit level. In most cases the speech units are phonemes or diphones. The selection of the utterances aims to cover as many units as possible in different phonetic and prosodic contexts in order to provide the necessary variability in the synthetic speech output [1]. Text to speech technology is now employed in a wide range of applications, spanning from assistive tools and education, to telecommunications and entertainment [2]-[5]. Application areas such as assistive aids and tools, speech-to-speech translation systems, robotics, mobile phones, household devices, navigation and personal guidance gadgets, can largely benefit from the more natural and intuitive means of human computer interaction (HCI) offered by speech [6]-[9].

Although recent statistical parametric approaches for speech synthesis such as Hidden Markov Models (aiming to model the features of one's voice through HMM), give promising results, the unit-selection approach provides a wide framework for speech synthesis, which has been advocated to fulfill most needs, as well as domains and computational environments [10]-[12]. Recent research in the speech synthesis field has been mainly concentrated on optimizing several aspects of the speech synthesis process, such as expressivity, advanced signal processing and more [2].

Although the speech quality achieved by today's general-purpose speech synthesis systems is considered to be quite high and closer to natural than ever, experience of the field indicates systems that are tailored to specific domains and application requirements, achieve higher quality results and performance. Depending on the application context, issues such as the typical trade-off between speech quality and synthesis speed and responsiveness, or processing requirements and storage space, can significantly deviate.

Especially in the case of deploying a speech synthesis system in the context of assistive technologies and tools, exploiting any available domain-specific *a priori* knowledge in order to identify and meet special requirements, can make a significant difference to the end-users [12]-[13]. In speech synthesis, as in most language technologies, language-dependent and language-independent issues need to be considered.

For example, speed optimization is a language-independent issue which calls for specific design choices at the algorithmic and data representation level, although the specific characteristics of certain languages may raise the need for special fine-tuning on a per-language basis. On the contrary, text normalization and efficient handling of particular language phenomena, are clearly language-dependent issues. In this paper, we describe design issues and implementation approaches for optimizing the performance of a general-purpose unit selection TTS technology for use into screen reading environments aimed to assist blind users. These special issues arise from the fact that when integrated into a screen reading environment, a TTS system is asked to

[1] A. Chalamandaris, S. Karabetsos, P. Tsiakoulis and S. Raptis are affiliated with the Institute for Language and Speech Processing (ILSP) / R.C. Athena, Department of Voice & Sound Technology and innoetics ltd, Artemidos 6 & Epidavrou, Marousi, GR 15125, Athens, Greece (e-mail: {achalam, sotoskar, ptsiak , spy}@ilsp.gr).

Contributed Paper
Manuscript received 07/14/10
Current version published 09/23/10
Electronic version published 09/30/10.

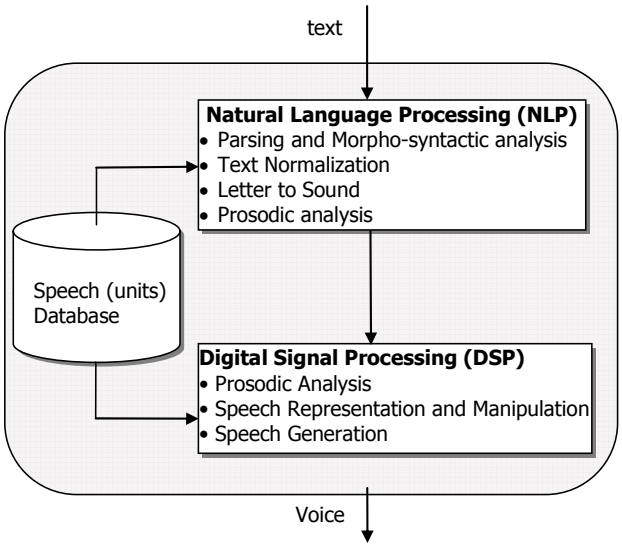

Fig. 1. General architectural diagram of a corpus-based TTS system.

synthesize many types of text input, such as chat dialogues, e-books, emails, automatically recognized scanned text by OCR systems, web pages etc. Along with this variability of inputs, blind users tend to use TTS systems in settings that sighted people most of the times find impossible to understand [14], while at the same time they pose requirements that are essential to fulfill, such as responsiveness, flexibility and intelligibility. These characteristics are crucial factors for the satisfaction or frustration of the end-user.

Keeping in mind all the above, emphasis is given in three main issues research has shown that need special care when integrating TTS technology into screen reading environments; namely natural language processing, speed and quality optimization.

The rest of this paper is organized as follows. In section II, the unit selection concatenative speech synthesis technology is briefly reviewed and a description of the TTS system's architecture is given, highlighting its core modules. Section III provides details on the followed design and implementation techniques related to the efficient integration of the TTS technology into screen reading environments. In section IV, the results of a subjective evaluation stage are presented regarding the performance of the delivered system in the framework of assistive tools. Finally a summary and some conclusive remarks are given in section V.

II. TEXT TO SPEECH SYSTEM ARCHITECTURE

The general architecture of a corpus-based TTS system is depicted in Fig. 1. There are two main components most often identified in such a system, namely the Natural Language Processing unit (NLP) and the Digital Signal Processing unit (DSP). This schematic applies for every data driven (i.e. any corpus-based) TTS system, regardless of the underlying technology (e.g., unit selection or parametric) [1]. The NLP component accounts for every aspect of the linguistic processing of the input text, whereas the DSP component accounts for the speech signal manipulation and the output generation. For a unit selection TTS, besides the speech units (usually diphones) the speech database contains all the necessary data for the unit selection stage of the synthesis [1], [5]. These components deserve more explanation.

In particular, the NLP component is mainly responsible for parsing, analyzing and transforming the input text into an intermediate symbolic format, appropriate to feed the DSP component. Furthermore, it provides all the essential information regarding prosody, that is, pitch contour, phoneme durations and intensity. It is usually composed of a text parser, a morpho-syntactic analyzer, a text normalizer, a letter-to-sound module [15] and a prosody generator. All these components are essential for disambiguating and expanding abbreviations and acronyms, for producing correct pronunciation, and also for identifying prosody related anchor points.

The DSP component includes all the essential modules for the proper manipulation of the speech signal, that is, prosodic analysis and modification, speech signal representation processing and generation. Among various algorithms for speech manipulation, *Time Domain Pitch Synchronous Overlap Add* (TD-PSOLA), *Harmonic plus Noise (HNM)*, *Linear Prediction based* (LPC-based) and *Multiband Resynthesis Overlap Add* (MBROLA) are the techniques that are mostly employed. Aside from the aforementioned modules, the DSP component includes also the unit selection module, which performs the selection of the speech units from the speech database using explicit matching criteria [1]. A more detailed architectural diagram of our TTS system is illustrated in Fig. 2.

A. NLP

As shown in Fig. 2, the input text is fed into the parsing module, where sentence boundaries are identified and extracted. This step is important since all of the following modules perform exclusively sentence-level text processing. The identified sentences are then fully expanded by the text normalization module. This expansion addresses numbers, abbreviations and acronyms as well as special tokens fused into the process by a screen reading software (e.g., chat dialogues, emails etc). For such cases, particular care must be taken for the proper manipulation and expansion of special strings such as abbreviated menu options, stress disambiguation and alternative ways of text writing (e.g., multilingual, misspelled, OCR misses or even transliterations-Romanization, like in the case of *greeklish* [16]). The term *greeklish* stands for a combination of the Greek and the English language (e.g., Greek-lish) and it consists of a transliterating manner of Greek text writing using the Latin alphabet. This Romanization is used frequently in e-mail communication among Greek-speaking computer users, and its main characteristic is the lack of a standardized table of transliteration mapping. In order to deal with these issues, the

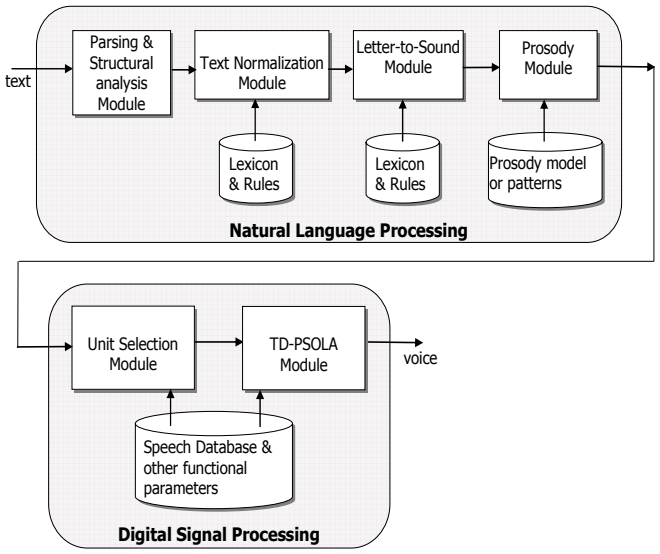

Fig. 2. System architecture of the unit selection TTS system.

text normalization module relies on a rule-based approach combined with lexicon resources. The letter-to-sound module transforms the normalized text into an intermediate symbolic form providing text's phonetic description. This module relies on a rule-based approach, complemented by exception dictionaries when necessary [15].

B. DSP

The DSP component comprises the unit selection and the signal manipulation modules, which in our case the latter is based on the TD-PSOLA algorithm. The speech database of the TTS system is encoded at a sampling frequency of 22 KHz and it includes annotated diphones as principal speech units, derived from the recordings of a Greek female professional speaker.

The unit selection module is considered to be one of the most important components in a corpus-based concatenative speech synthesis system. It provides a mechanism to automatically select the optimal sequence of database units that produce the final speech output, the quality of which depends on its efficiency. The criterion for optimizing is the minimization of a total cost function which is defined by two partial cost functions, namely the target cost and the concatenation cost functions [1], [5].

The target cost function measures the similarity of an applicant unit with its predicted specifications (from NLP) and is defined as,

$$C^t(t_i,u_i)=\sum_{j=1}^{p} w_j^t \cdot C_j^t(t_i,u_i) \tag{1}$$

where, $u_1^n=\{u_1, u_2, \ldots, u_n\}$ are the candidate (sequence) units, $t_1^n=\{t_1, t_2, \ldots, t_n\}$ are the target (sequence) units, $C_j^t(t_i,u_i)$ is a partial target cost, p is the dimension of the target feature vector and w_j^t is a weighting factor for every partial target cost. The target feature vector typically employs target values for prosody and contextual features. The concatenation (or aka join) cost function accounts for the acoustic matching between pairs of candidate units and is defined as,

$$C^c(u_{i-1},u_i)=\sum_{j=1}^{q} w_j^c \cdot C_j^c(u_{i-1},u_i) \tag{2}$$

where, $C_j^c(u_{i-1},u_i)$ is a partial join cost , q is the dimension of the join feature vector and w_j^c is a weighting factor for every partial join cost. The feature vector typically includes similarity measurements for the spectral, pitch and contextual dimensions. Hence, the total cost is defined as,

$$C(t_1^n,u_1^n)=\sum_{i=1}^{n} W^t \cdot C^t(t_i,u_i)+\sum_{i=2}^{n} W^c \cdot C^c(u_{i-1},u_i) \tag{3}$$

or based on (1) and (2) it can be written as,

$$C(t_1^n,u_1^n)=\sum_{i=1}^{n} W^t \cdot \sum_{j=1}^{p} w_j^t \cdot C_j^t(t_i,u_i)+\sum_{i=2}^{n} W^c \cdot \sum_{j=1}^{q} w_j^c \cdot C_j^c(u_{i-1},u_i) \tag{4}$$

where, W^t and W^c are the weights that denote the significance of the target and the join costs, respectively. The goal of the unit selection module is to perform a (computationally demanding) search, so as to find the speech unit sequence which minimizes the total cost, hence to specify,

$$\hat{u}_1^n=\min_{u_1 \ldots u_n} C(t_1^n,u_1^n) \tag{5}$$

The selection of the optimal speech unit sequence incorporates a thorough search (usually a Viterbi search) which involves comparisons and calculations of similarity measures between all available units, often employing heuristics to guide and/or limit the search [1], [5] for higher efficiency.

III. ADAPTATION ISSUES FOR DESIGNING AN ASSISTIVE TOOL

Although TTS technology in general offers high quality synthetic voice, it needs to be specially adapted and customized for dedicated services or tools. This is made clear if one considers the different user requirements that arise from the application context. A telecom application would pose different requirements to be met, than educational software would. Similarly, in the case of an assistive tool for blind people, the design requirements are significantly different from other cases. As already pointed out, the TTS component of a screen reading platform is the last stage of the interface between the computer and the user, it is desired to be able to cope with almost all kinds of text. At the same time, studies and experience have shown that blind computer users prefer extreme settings in the TTS system, with minimal response delay. Furthermore, in the ideal case, the users would prefer to use the same voice for reading aloud all possible texts, even if they are in different language, without having to compromise as far as the voice quality is concerned. All these issues are addressed in the following paragraphs along with the approach we adopted for better results.

A. NLP Module Adaptation

The adapted NLP module shown in Fig. 3, is responsible for the parsing, analysis and processing of the input text, in order for the DSP module to be able to produce the synthetic

speech output. In our case, it was necessary to design an NLP component that would be able to cope with text, mainly written in Greek, but also incorporating non-Greek or *greeklish* [16] textual segments. Dealing with *greeklish* efficiently is not a trivial issue, especially when the text is asked to be pronounced by a TTS system. Previous research of the authors [16] has resulted in the development of a robust and accurate technology for identifying and converting *greeklish* words and phrases into correct Greek. This technology incorporates an intermediate stage of language identification based on a Bayesian acoustic model for the Greek language. The input text is firstly identified and categorized as *greeklish* and non-*greeklish* chunks and processed accordingly. We need to note here however, that although the above transliteration technology is very accurate, it was decided at a later stage of the development to provide the user with the option to enable and disable this feature in the NLP module, depending on the input text source (e.g. websites, documents, chat dialogues etc).

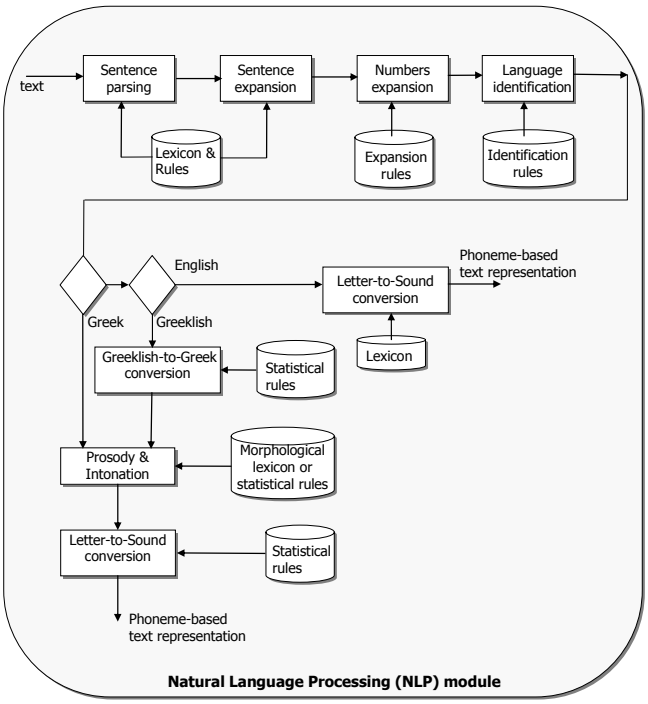

Fig. 3. The NLP module utilized in the unit selection TTS system for screen readers.

B. Speed Optimization and Response Latency

Speed optimization and response latency minimization are two issues that blind users identified as crucial in our research. It is important for a blind user to be able to listen to synthetic speech almost immediately after he/she pushes a button on the keyboard; otherwise the user gets frustrated if a noticeable delay between action and response is present. It was to our surprise that many users suggested to include the option of allowing for degraded speech quality in exchange for increased speed. Various previous surveys report on typical use patterns and configuration settings employed by blind users when working with screen-readers [17].

In order to meet these requirements, two strategies were followed: a) a database size reduction and b) pre-recording and pre-synthesis of essential for the task words and phrases. Previous research on these topics has resulted in efficient techniques for database coding and compression, as well as database reduction and speed optimization for TTS systems with minimal speech quality degradation (e.g., [18]-[23]). Although previous work of the authors aimed to reduce the speech database into significantly smaller size for incorporation into portable devices, the same method was also followed in this case, with looser restrictions; namely, from an initial 8-hour long database, we derived a 4-hour long subset with comparable performance and which was considered fully acceptable for the specific application context. More details about our database reduction approach can be found in [22], [23].

In order to minimize system's response latency and to maintain high speech quality when the user is typing, we incorporated into our database all the recordings which are necessary for pronouncing all possible keyboard strokes. In other words, all Latin and Greek letter names, numbers and punctuation marks were separately recorded by the same speaker and incorporated into our TTS database. Additionally, in our DSP module, we introduced preselected Viterbi paths for all these utterances [19], in order to minimize synthesis processing time for such cases. By doing so, we achieved high-quality and fully intelligible synthetic speech for instant keyboard typing feedback, allowing at the same time the DSP module to fully manipulate the signal and the user to set the speech intonation and rate at his request and preference.

C. Multilingualism and Performance

One of the most important factors of blind users' satisfaction was noted to be the ability of the TTS system to pronounce both Greek and English text with the same voice and without degradation of the intelligibility. Most of the already available non-English TTS systems, whenever they are asked to pronounce an English word or phrase, they either map the orthographic representation of the text into system's native language, or they perform a letter to sound conversion for the English language and then force a mapping between the English and the native phoneme set. This approach succeeds in pronouncing embedded English words, but with very low intelligibility, noticeable inconsistency and low usability.

In order to deal with this issue, our strategy was to create an entirely new spoken corpus for English with the same speaker; in other words we created a small-scale but fully functional native English TTS system with the same speaker. This method ensures better intelligibility in English, since the TTS system employs the appropriate phoneme set for the English language, while at the same time it incorporates an English text-processing unit. It is obvious that our strategy demands extra effort and development; nevertheless, our experiments

asserted that this is a crucial advantage, as far as a blind user's satisfaction is concerned.

By adopting this approach there were extra issues that needed to be addressed, such as the implementation of the appropriate letter to sound module, the design and the recording of an additional database and many more. Among those, the most important one was to ensure that the letter to sound module was properly customized to the specifics of the speaker. In other words, since the speaker we used for the Greek database was not a native English speaker, any variations in his English pronunciation and accents should be appropriately annotated and fused into the respective letter to sound module we would incorporate into our TTS system. Keeping all the above in mind one can safely deduce that the knowledge of English was set to be a crucial factor for selecting the original speaker for our recordings.

IV. Evaluation and Results

The techniques described in this work are assessed using subjective criteria since the task is to meet the target user requirements. For subjective evaluation, the most common approach for assessing the quality of TTS systems is through listening tests where a group of people is asked to express their opinion regarding the TTS quality. The TTS quality refers mainly to the produced synthetic speech and it is usually defined in terms of naturalness and intelligibility but it sometimes also addresses other dimensions of speech perception such as, acceptability, comprehension, impression, pronunciation, pleasantness, listening effort and other [24], [25]. The listening tests aim mainly to assess the aspects of naturalness and intelligibility by employing word-level, sentence-level, as well as paragraph-level evaluation. The results are expressed in terms of mean opinion scores (MOS), reflecting rather accurately the perceived quality of the synthetic speech output of a TTS system [1], [5], [24], [25]. In our case however, where the developed system was necessary to address specific needs of the target users, a usability evaluation was also performed in order to assess the integration of the resulted systems in screen reading environments [26].

The experiments were designed using our TTS system and a bilingual speech database (Greek and English) from the same speaker. After optimization, the Greek database was 4-hour long, while the English one was 48-minutes long, both containing more than 300k instances of the covered diphones.

A. Speech Quality Assessment

To assess the quality of the speech synthesis system, a set of acoustic experiments was performed. As already mentioned, the experiments aimed to evaluate different dimensions of the synthetic speech quality via sentence-level, word-level and paragraph-level acoustic tests. The test subjects were 15 native Greek-speaking people with visual impairements (either blind or partially impaired vision), who already had significant experience with TTS systems. Normally the test subjects in such evaluation tests receive a short training session in order to get accustomed to the nature of the synthetic speech. However in our case, the test subjects have already long experience with synthetic speech and therefore they did not require any training in order to complete the experiments successfully.

1) Experiment 1: Sentence-level evaluation

The aim of the first experiment was to evaluate the performance of the TTS system in terms of naturalness (i.e. how close to natural the synthetic speech is), ease of listening (i.e the effort that is necessary in order to follow and understand what is being said) and articulation (whether the speech is clearly articulated). The Mean Opinion Score (MOS) was used as the subjective scoring method. The stimuli consisted of 35 randomly selected, medium-sized sentences with an average of 13 words per sentence. The sentences where synthesized using the derived text-to-speech system, and the listeners were asked to rate the three aforementioned quality dimensions for each sentence by grading on a scale of 1 to 5 for each dimension. In order to ensure consistency in the responses, each grade was assigned a label. This is shown in table I, while table II summarizes the mean scores (MOS) and the standard deviations of the responses. It is worth noting that the "ease of listening" and the "articulation" received remarkably high grades. Furthermore, the overall score for "naturalness" which lies near 4 is particularly high, considering that 4 corresponded to "near natural". It is worth noting here that the results for "ease of listening" of the speech output are quite high, illustrating that the synthetic speech contains minimal number of distractions that would otherwise demand more effort from the listener for perceiving the transmitted message.

TABLE I
Scale Labels for MOS Evaluation: Experiment 1

	Naturalness	Ease of listening	Articulation
1	Unnatural	No meaning understood	Bad
2	Inadequately natural	Effort required	Not very clear
3	Adequately natural	Moderate effort	Fairly clear
4	Near natural	No appreciable effort required	Clear enough
5	Natural	No effort required	Very clear

TABLE II
Evaluation Results: Experiment 1

	Naturalness	Ease of listening	Articulation
MOS	3.72	4.45	4.15
STD	0.68	0.70	0.72

2) Experiment 2: Word-level intelligibility evaluation

The aim of this phonetic task was to evaluate the TTS system in terms of intelligibility. In order to do so, the Diagnostic Rhyme Test (DRT) was employed, which provides a widely

used index for diagnostic and comparative evaluation of the intelligibility of single initial or final consonants. The stimuli used consisted of 33 groups of two or three words each, some of which were nonsense. The words in each group were only differentiated in one letter. For each group, the participants were presented with the list of words and one of them was synthesized and played back. They were then asked to select which word from the list they heard. In the vast majority of cases, namely over 98.3%, all participants were able to correctly match the stimulus with the respective word in the list.

TABLE III
SCALE LABELS FOR MOS EVALUATION: EXPERIMENT 3

	Naturalness	Ease of listening	Pleasantness	Intelligibility	Pronunciation
1	Bad	No meaning understood	Very unpleasant	Unclear all the time	Very frequent pronunciation irregularities
2	Poor	Effort required	Unpleasant	Not very clear	Frequent pronunciation irregularities
3	Fair	Moderate effort	Fair	Fairly clear	Few pronunciation irregularities
4	Good	No appreciable effort required	Pleasant	Clear enough	Rarely any pronunciation irregularities
5	Excellent	No effort required	Very Pleasant	Very clear	No pronunciation irregularities at all

TABLE IV
EVALUATION RESULTS: EXPERIMENT 3

	Naturalness	Ease of listening	Pleasantness	Intelligibility	Pronunciation
MOS	3.62	3.72	3.70	3.77	3.50
STD	0.78	0.85	0.88	0.72	0.79

3) Experiment 3: Paragraph-level evaluation

The aim of this task was to evaluate TTS system's quality in terms of speech flow and to obtain feedback on the overall listening experience as perceived at a level higher than a single sentence. Several aspects of the synthetic speech quality were addressed; hence, the participants were asked to evaluate both overall achieved naturalness (that is, how natural does the synthetic speech sound if compared to human reading) as well as particular aspects regarding, a) ease of listening (effort to follow and understand what is being said), b) pleasantness (if the synthetic voice is pleasant to hear), c) intelligibility (if words and phrases were well understood) and d) pronunciation (appearance of pronunciation irregularities). The listeners were asked to grade the above aspects of the stimuli on a 1 to 5 scale (MOS assessment). Again each grade was assigned appropriate labels, as shown in table III, in order to achieve consistency among people. The MOS results are summarized in table IV. The stimuli consisted of 5 randomly selected paragraphs with an average of 6 sentences (or 83 words) per paragraph.

The evaluation results show that the TTS system achieves good performance as far as the aforementioned dimensions are concerned, providing a good illustration of the overall achieved quality [25].

B. Usability Evaluation

Usability evaluation is an important link to the iterative design process, especially in our case where the application context of the developed system presents specific needs as far as the target group is concerned [26]. In order to evaluate the usability of the system a two-phase small-scale process was carried out. The process consisted of a heuristic and a laboratory phase. During the heuristic evaluation we attempted to identify issues that accessibility experts would regard as obstacles from the end-user's point of view. The second phase of the evaluation process, which included the evaluation by potential target users, blind users assessed the three main aspects of the system, namely effectiveness, efficiency and satisfaction.

1) Heuristic evaluation phase

During the heuristic evaluation phase, the evaluators, three experts in accessibility issues with great experience in design and evaluation of accessible software, performed different tasks in a screen reading environment, without the use of a monitor display. The subjects were given specific tasks to fulfill, while they were asked to provide feedback about their satisfaction level while using the TTS system in combination with the screen reading environment. This evaluation phase acted complementarily to the actual experimental phase where target users were employed to thoroughly assess the system. The results from this phase were mainly used during the system's iterative design method we adopted for developing the TTS system.

2) Experimental evaluation phase

The actual laboratory experimental evaluation phase was carried out with the help of 6 blind computer users with significant expertise in computer interaction through screen reading environments. This phase consisted of three one-hour sessions for each subject, where the participants were asked to fulfill several tasks such as web browsing, book reading and more. The results from every session were collected manually through interviews and through observation of the subjects while using the system. Although the number of the evaluators was not large enough, it was very helpful for reaching to important conclusions about the usability of our system when integrated into screen reading environments.

a) Effectiveness

All subjects managed to complete all the tasks that had been given, without any problem. They noted that advantageous features such as the advanced NLP module and the use of the same voice for pronouncing English words and phrases was very effective in their interaction with the screen reading platform. Tasks that normally would be difficult or even impossible to complete otherwise, such as email or chatting reading in *greeklish* were now possible and easy to complete.

b) Efficiency

All subjects provided positive feedback about the efficiency of the system when integrated into a screen reading environment. The responsiveness, the clear voice in voice-assisted typing and the high intelligibility even in very fast speed rates were the main issues that were identified as the most positive aspects of system's performance.

c) Satisfaction

As far as overall satisfaction was concerned, all evaluators answered overly satisfied by the features the system provided and its performance in combination with screen reading software. Special focus was given in system's advanced NLP features.

V. CONCLUSIONS

In this paper, we have described the system architecture and elaborated on advanced features and functionalities of a unit selection TTS system that is designed to effectively collaborate with screen reading software as an assistive tool for visually impaired computer users. We presented techniques commonly used in similar systems, along with our strategies and approaches whenever they differentiate. Transliteration issues such as *greeklish* were addressed efficiently with the use of a statistically driven transliteration technology for converting *greeklish* to Greek. Finally, advanced performance issues, such as low latency and low computational complexity were also addressed with a reduced in size database and with the incorporation of prerecorded utterances for assisted typing. Evaluation results provide clear evidence of the high performance of our TTS system, as far as its output quality is concerned, but also as far as its performance when integrated into a screen reading environment. Currently the system has been adopted by the National Association for Blind in Greece.

ACKNOWLEDGMENT

The authors would like to thank National Association for Blind in Greece, for its important assistance during evaluation phase, Mr. C. Kouvaris, Mr. G. Simeonides, Mr. L. Papadopoulos Mrs. E. Carypedou, Mr. L. Chrysses and Mr. M. Alexandrakis for their contribution during user requirements design stage, as well as during the beta testing evaluation phase.

REFERENCES

[1] T. Dutoit, "Corpus-based Speech Synthesis*," Springer Handbook of Speech Processing,* J. Benesty, M. M. Sondhi, Y. Huang (eds), Part D, Chapter 21, pp. 437-455, Springer, 2008.

[2] G. Bailly, W.N. Campbell, and B. Mobius, "ISCA Special Session: hot topics in speech synthesis", *Proc. Eurospeech 2003,* pp. 37-40, Geneva, 2003.

[3] B. Duggan and M. Deegan, "Considerations in the usage of text to speech (tts) in the creation of natural sounding voice enabled web systems", *In Proc. of the 1st Int. Symp. on Information and Communication technologies (ISICT '03:)*, pp. 433–438, Trinity College Dublin, 2003.

[4] N. Campbell, "Developments in Corpus-Based Speech Synthesis: Approaching Natural Conversational Speech," *IEICE trans. Inf. & Syst.*, vol. E88-D, no. 3, pp.376-383, 2005.

[5] Douglas O'Shaughnessy, "Modern Methods of Speech Synthesis," *IEEE Circuits and Systems Magazine,* Third Quarter 2007, pp. 6-23, 2007.

[6] T. Schultz, A. W. Black, S. Vogel, and M. Woszczyna, "Flexible Speech Translation Systems," *IEEE trans. on Audio, Speech and Language Processing,* vol. 14, no. 2, pp. 403-411, 2006.

[7] S. Tomko, T. K. Harris, A. Toth, J. Sanders, A. Rudnicky and R. Rosenfeld, "Toward Efficient Human Machine Speech Communication: The Speech Graffiti Project," *ACM Trans. on Speech and Language Processing*, vol. 2, no. 1, Article 2, pp. 1-27, 2005.

[8] R. K. Moore, "PRESENCE: A human-inspired architecture for speech-based human-machine interaction," *IEEE Trans. Computers*, 56, pp. 1176-1188, 2007.

[9] L. Mohasi and D. Mashao, "Text-to-Speech Technology in Human-Computer Interaction", *5th Conference on Human Computer Interaction in Southern Africa, (CHISA 2006, ACM SIGHI)*, pp. 79-84, 2006.

[10] M. Schnell, O. Jokisch, R. Hoffmann, and M. Kustner, "Text-to-speech for low-resource systems," *IEEE Workshop Multimedia Signal Processing (MMSP)*, St. Thomas, pp. 259-262, 2002.

[11] Kim S.-J., Kim J.-J. and Hahn M.-S., "HMM-based Korean speech synthesis system for hand-held devices," *IEEE Trans. Consumer Electronics*, vol. 52, no. 4, pp. 1384-1390, 2006.

[12] Earl, C.L., Leventhal, J.D., "A survey of windows screen reader users: Recent improvements in accessibility," *Journal of Visual Impairment and Blindness*, vol. 93, no. 3, pp. 174-177, 1999.

[13] J. Bigham, Cavender A. C., J. T. Brudvik, J. O. Wobbrock, and R. Ladner, "WebinSitu: A comparative analysis of blind and sighted browsing behaviour," *9th Intl. Conf. ACM SIGACCESS on Computers and Accessibility*, Arizona USA, pp. 51-58, 2007.

[14] K. Barnicle, "Usability testing with screen reading technology in a windows environment," *Proc. of the Conf. on Universal Usability*, pp. 102-109, 2000.

[15] A. Chalamandaris, S. Raptis, and P. Tsiakoulis, "Rule-based grapheme-to-phoneme method for the Greek," *in Interspeech 2005*, pp. 2937-2940, 2005.

[16] A. Chalamandaris, A. Protopapas, P. Tsiakoulis, and S. Raptis. "All Greek to me! An automatic Greeklish to Greek transliteration system." *5th Int. Conf. on Language Resources and Evaluation (LREC 2006)*. Genoa, Italy, pp. 1226–1229, 2006.

[17] G. Redish, and Theofanos, M.F, "Observing users who listen to web sites," *Usability Interface*, vol.9, no.4, 2003.

[18] Chu, Wai C. "*Speech coding algorithms: Foundation and evolution of standardized coders,*" John Wiley & Sons, 2003.

[19] M. Beutnagel, Mohri, R., and Riley, M., "Rapid unit selection from a large speech corpus for concatenative speech synthesis," *in Proc. Eurospeech 99*, Budapest, 1999.

[20] A. Black, and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," *Proc. of Eurospeech 97*, vol. 2, pp. 601-604, Greece, 1997.

[21] G. Coorman, Fackrell, J., Rutten, P., and Coile, B. V., "Segment selection in the LH realspeak laboratory TTS system," *Proc. of the ICSLP 2000*, vol. 2, pp. 395-398, 2000.

[22] P. Tsiakoulis, A. Chalamandaris, S. Karabetsos and S. Raptis, "A statistical method for database reduction for embedded unit selection speech synthesis," *in IEEE ICASSP 2008*, pp. 4601-4604, 2008.

[23] S. Karabetsos, P. Tsiakoulis, A. Chalamandaris, and S. Raptis, "Embedded unit selection text-to-speech synthesis for mobile devices," *IEEE Trans. on Consumer Electronics*, vol. 55, no. 2, pp. 613-621, 2009.

[24] V. J. van Heuven and R. van Bezooijen, "Quality Evaluation of Synthesized Speech," *Speech Coding and Synthesis,* W. B. Kleijn, and K. K. Paliwal (eds), Chapter 21, pp. 707-738, Elsevier Science, 1995.

[25] M. Viswanathan and M. Viswanathan, "Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale," *Computer Speech & Language*, vol. 19, pp. 55-83, January 2005.

[26] J. T. Hackos and J. C. Redish, "*User and task analysis for interface design*," John Wiley & Sons, Inc., Chichester 1997.

BIOGRAPHIES

Aimilios Chalamandaris received his M. Eng. degree in Electrical Engineering and Computer Science from the National Technical University of Athens in 2000, and his M. Eng in Telecoms and Signal Processing from Imperial College in 2001. He is currently working towards his PhD at the National Technological University in Athens and he is working at the Institute for Language and Speech Processing (ILSP) – Athena Research Centre, doing research on speech and signal processing and affiliates with innoetics ltd. His research interests are speech synthesis, speech recognition, NLP, and signal processing.

Sotiris Karabetsos received the M. Eng. degree in Electrical Engineering and Computer Science from the National Technical University of Athens (NTUA), in 2004 and the M.S. degree in Data Communications from Brunel University of London, in 2003. He has also received the BS degree in Electronic Engineering from the Technological and Educational Institution of Athens (TEI of Athens), in 1999. From 2003, he is with the Institute for Language and Speech Processing (ILSP). He also affiliates with innoetics ltd. He is also with the Technological and Educational Institution of Athens (TEI of Athens), Department of Electronics. His research interests are speech synthesis, signal processing, and telecommunications. He is an IEEE member.

Pirros Tsiakoulis received his M. Eng. degree in Electrical Engineering and Computer Science in 2003 from the National Technical University of Athens (NTUA), Athens, Greece. He received his Ph.D. from National Technical University of Athens (NTUA) in 2010. In 2000, he joined the Institute for Language and Speech Processing (ILSP). He is also affiliated with innoetics ltd. His research interests include speech synthesis, NLP, speech recognition and speech processing. He is a member of IEEE.

Spyros Raptis received his M. Eng. degree in Electrical Engineering and Computer Science in 1994 and his PhD in hybrid computational intelligence for optimization, modeling and decision making in 2001, both from the National Technical University of Athens, Greece. He has been a lecturer at graduate and post-graduate level and has participated in a number of National and European RTD projects on speech technology, computational intelligence, robotics, and multimedia educational applications. He is currently a researcher at the Voice & Sound Technology Department at the Institute for Language and Speech Processing (ILSP) leading the speech synthesis team. He also affiliates with innoetics ltd. His research interests include speech processing and applications, computational intelligence, software agents, hybrid systems and robotics.